FEMS Microbiology Letters, 365, 2018, fny120

doi: 10.1093/femsle/fny120
Advance Access Publication Date: 12 June 2018
Research Letter

RESEARCH LETTER – Pathogens & Pathogenicity

Genomic characterisation of an international *Pseudomonas aeruginosa* reference panel indicates that the two major groups draw upon distinct mobile gene pools

Luca Freschi[1,†], Claire Bertelli[2,3,†], Julie Jeukens[1], Matthew P. Moore[4], Irena Kukavica-Ibrulj[1], Jean-Guillaume Emond-Rheault[1], Jérémie Hamel[1], Joanne L. Fothergill[3], Nicholas P. Tucker[5], Siobhán McClean[6], Jens Klockgether[7], Anthony de Soyza[8], Fiona S.L. Brinkman[2,†], Roger C. Levesque[1,†] and Craig Winstanley[4,*,†,‡]

[1]Institute for Integrative and Systems Biology (IBIS), University Laval, Québec City, QC G1V 0A6, Canada, [2]Department of Molecular Biology and Biochemistry, Simon Fraser University, Burnaby, BC V5A 1S6, Canada, [3]Institute of Microbiology, University Hospital Center and University of Lausanne, CH-1011 Lausanne, Switzerland, [4]Institute of Infection and Global Health, University of Liverpool, Liverpool L69 7BE, UK, [5]Strathclyde Institute of Pharmacy and Biomedical Sciences, University of Strathclyde, Glasgow G4 0RE, UK, [6]Centre of Microbial Host Interactions, Institute of Technology Tallaght, Tallaght, Dublin D24 FKT9, Ireland, [7]Clinic for Paediatric Pneumology, Allergology, and Neonatology, Hannover Medical School, D-30625, Hannover, Germany and [8]Institute for Cellular Medicine, Newcastle University, Newcastle-upon-Tyne NE2 4HH, UK

*Corresponding author: Institute of Infection and Global Health, University of Liverpool, Ronald Ross Building, 8 West Derby Street, Liverpool L69 7BE, UK. Tel: +44 0151 795 9642; Fax: +44 0151 795 5527; E-mail: C.Winstanley@liv.ac.uk
†These authors contributed equally to this work.
One sentence summary: By genome sequencing *Pseudomonas aeruginosa* from a reference panel, we show that the organism divides into two major groups on the basis of both SNP phylogeny and accessory genome content.
Editor: Kendra Rumbaugh
‡Craig Winstanley, http://orcid.org/0000-0002-2662-8053

ABSTRACT

Pseudomonas aeruginosa is an important opportunistic pathogen, especially in the context of infections of cystic fibrosis (CF). In order to facilitate coordinated study of this pathogen, an international reference panel of *P. aeruginosa* isolates was assembled. Here we report the genome sequencing and analysis of 33 of these isolates and 7 reference genomes to further characterise this panel. Core genome single nucleotide variant phylogeny demonstrated that the panel strains are widely distributed amongst the *P. aeruginosa* population. Common loss-of-function mutations reported as adaptive during CF (such as in *mucA* and *mexA*) were identified amongst isolates from chronic respiratory infections. From the 40 strains analysed,

Received: 5 January 2018; **Accepted:** 14 May 2018

37 unique resistomes were predicted, based on the Resistance Gene Identifier method using the Comprehensive Antibiotic Resistance Database. Notably, hierarchical clustering and phylogenetic reconstructions based on the presence/absence of genomic islands (GIs), prophages and other regions of genome plasticity (RGPs) supported the subdivision of *P. aeruginosa* into two main groups. This is the largest, most diverse analysis of GIs and associated RGPs to date, and the results suggest that, at least at the largest clade grouping level (group 1 vs group 2), each group may be drawing upon distinct mobile gene pools.

Keywords: *Pseudomonas aeruginosa*; comparative genomics; antimicrobial resistance; genomic islands

INTRODUCTION

Pseudomonas aeruginosa is a leading cause of nosocomial and other opportunistic infections, especially in relation to chronic lung infections of patients with the genetically inherited disease cystic fibrosis (CF) (Lyczak, Cannon and Pier 2000; Cohen and Prince 2012). Increasingly, it is associated with high levels of multidrug resistance, with important clinical and economic consequences (Nathwani *et al.* 2014). Indeed, *P. aeruginosa* has been included in the group of bacteria (the ESKAPE pathogens) most associated with the worrying increases in antimicrobial resistance (AMR) (Pendleton, Gorman and Gilmore 2013) and has been identified by the World Health Organization as one of the top three priority pathogens urgently requiring new antimicrobial therapies for treatment.

Much of the research carried out into the mechanisms of virulence of *P. aeruginosa* has been focused on a limited number of strains, most notably strain PAO1, which many consider to be a laboratory strain, and which has itself diversified during its existence in multiple laboratories (Stover *et al.* 2000; Klockgether *et al.* 2010). Taking into account the diversity in phenotypic behaviour and population structure within the species (Freschi *et al.* 2015), and the desirability of using relevant clinical isolates, a strain panel of diverse *P. aeruginosa* strains was assembled (De Soyza *et al.* 2013). The panel was chosen to represent diversity in source (clinical, environmental, and geographical) and phenotype. Subsequently, detailed phenotypic characterisation was carried out in order to clearly define the characteristics of the panel strains (Cullen *et al.* 2015).

The global *P. aeruginosa* population is highly diverse, but also contains some abundant clones, such as the PA14-like lineage and Clone C (Cramer *et al.* 2012; Hilker *et al.* 2015). Since the publication of the first complete *P. aeruginosa* genome sequence in 2000 (Stover *et al.* 2000), there has been considerable progress with the comparative genomics of the species, with a number of studies reporting analyses of multiple genomes (Mathee *et al.* 2008; Jeukens *et al.* 2014; Stewart *et al.* 2014; Kos *et al.* 2015; van Belkum *et al.* 2015). Other studies have focused on genomic variations within individual lineages (Williams *et al.* 2015; Fischer *et al.* 2016). As well as helping us to resolve the phylogeny of *P. aeruginosa*, these studies have revealed key genomic features that vary between strains and contribute to the diversity of the species, including the islands and prophages that dominate the accessory genome (Pohl *et al.* 2014).

Adaptation and phenotypic diversification are key features of long-term chronic lung infections in CF patients (Winstanley, O'Brien and Brockhurst 2016), emphasising the difficulty in inferring mechanisms of behaviour during infection on the basis of single isolates or strains. Hence, it is important to access a diverse panel of *P. aeruginosa* strains that can better represent the diversity. The International Pseudomonas aeruginosa Consortium was formed with the aim of genome sequencing >1000 *P. aeruginosa* genomes and constructing an analysis pipeline for the study of *P. aeruginosa* evolution, virulence and antibiotic resistance (Freschi *et al.* 2015). Here, as part of this larger endeavour, in order to better define the characteristics of the international *P. aeruginosa* reference panel of strains, we present comparative genomics analyses based on whole genome sequence data.

METHODS

Bacterial strains and growth conditions

The isolates used in this study are listed in Table 1. Bacterial colonies were isolated on DifcoTM Pseudomonas Isolation Agar (BD, Sparks, MD, USA). Strain NN1 from the original panel was omitted from this study because of contamination issues. Strains AA43 and AA44 were omitted at the request of the original suppliers of these isolates.

DNA extraction, library prep and genome sequencing

Genomic DNA was extracted from overnight cultures using the DNeasy Blood and Tissue Kit (QIAGEN, Hilden, Germany). Genomic DNA (500 ng) was mechanically fragmented for 40 s using a Covaris M220 (Covaris, Woburn, MA, USA) with default settings. Fragmented DNA was transferred to a tube and library synthesis was performed with the Kapa Hyperprep kit (Kapa Biosystems, Wilmington MA, USA) according to manufacturer's instructions. TruSeq HT adapters (Illumina, SanDiego CA, USA) were used to barcode the libraries, which were each sequenced in 1/48 of an Illumina MiSeq 300 bp paired-end run at the Plateforme d'Analyses Génomiques of the Institut de Biologie Intégrative et des Systèmes (Laval University, Quebec, Canada). Each dataset was assembled *de novo* with the A5 pipeline version A5-miseq 20140521 (Tritt *et al.* 2012). Where necessary, we resequenced some strains for which genome sequence data were already available. This was done to ensure uniform, higher quality genomes across the panel.

Core genome phylogeny

We performed a core genome phylogeny using the Harvest suite version v1.1.2 (Treangen *et al.* 2014). In addition to the panel strains, we included all strains present on NCBI for which an assembly with less than 30 scaffolds was available on November 2015.

Variant calling

For 38 panel strains (for which high-quality short read data were available), sequence reads were mapped to the genome of *P. aeruginosa* (PAO1) using the Burroes-Wheeler Alignment

Table 1. Summary of strains and genome sequence data.

Strain	Source	This study	Accession number	Genome size(bp)	N50 (bp)	Scaff (n)	Median coverage	acs	aro	gua	mut	nuo	pps	trp	MLST
LESB58	CF	N	NC_011770					6	5	11	3	4	23	1	146
LES400	CF	N	NZ_CP006982					6	5	11	3	4	23	1	146
LES431	CF	N	NC_023066					6	5	11	3	4	23	1	146
C3719	CF	Y	MCMM00000000	6192 913	409 151	31	35	28	5	11	18	4	13	3	217
DK2	CF	N	NC_018080												
AES-1R	CF	Y	MCML00000000	6343 337	414 654	32	27	11	84	11	3	4	4	7	649
AUS23 (AUST-02)	CF	Y	MCMN00000000	6272 404	485 799	44	57	28	5	11	5	4	4	7	775
AUS52	CF	Y	MCMK00000000	6209 179	963 855	23	34	28	5	5	11	3	15	44	242
AA2	CF	Y	MCMJ00000000	6258 177	371 531	50	20	11	3	11	3	1	4	60	708
AMT 0023–30	CF	Y	MCMI00000000	6471 685	478 937	31	25	11	5	6	3	74	13	7	1394
AMT 0023–34	CF	Y	MCMH00000000	6282 816	433 349	25	27	11	5	6	3	74	13	7	1394
AMT 0060–1	CF	Y	MCNB00000000	7036 907	260 085	75	79	17	5	5	4	4	4	3	111
AMT 0060–2	CF	Y	MCNA00000000	7037 467	302 236	75	89	17	5	5	4	4	4	3	111
AMT 0060–3	CF	Y	MCMZ00000000	7033 865	295 479	70	80	17	5	5	4	4	4	3	111
PAO1 (ATCC15692)	wound	N	NC_002516					7	5	12	3	4	1	7	549
UCBPP-PA14	burn	N	NC_008463					4	4	16	12	1	6	3	253
PAK	Non-CF	Y	MCMY00000000	6384 788	665 433	24	83	11	5	11	11	4	4	14	693
CHA	CF	Y	MCMG00000000	6512 494	495 272	25	56	14	5	10	155	4	13	7	1919
IST27	CF	Y	MCMX00000000	6332 447	574 655	29	94	4	4	3	–	1	6	–	–
IST27N	mutant	Y	MCMW00000000	6332 479	612 142	24	77	4	4	3	–	1	6	–	–
968333S	Non-CF Br	Y	MCMV00000000	6513 472	395 054	39	79	1	5	11	3	4	10	3	234
679	urine	Y	MCMU00000000	6410 839	371 307	39	76	11	5	11	11	3	27	7	198
39 016	keratitis	Y	MCLX00000000	6824 311	235 815	58	25	38	11	3	13	1	2	4	235
2192	CF	N	NZ_CH482384												NK
NH57388A	CF	Y	MCMT00000000	6197 427	415 910	28	63	28	5	11	11	4	12	3	387
1709–12	CF	Y	LZQH00000000	7129 475	288 354	56	86	17	5	5	4	4	4	3	111
Mi162	burn	Y	MCMF00000000	6586 986	198 657	109	22	13	4	5	5	12	7	15	308
Jpn1563	lake water	Y	MCMS00000000	6374 250	507 716	33	71	16	5	36	3	83	90	1	876
LMG 14 084	water	Y	MCMR00000000	7007 718	260 355	70	67	13	8	9	3	1	6	9	316
Pr335	HE	Y	MCMQ00000000	6679 225	539 576	45	96	6	5	6	7	4	6	7	27
U018a	CF	Y	MCMP00000000	6370 823	352 039	47	81	11	8	6	115	4	13	18	852
CPHL9433	tob. plant	Y	MCMO00000000	6421 125	674 344	38	86	27	13	9	156	1	7	192	1920
RP1	CF	Y	LNBU00000000	6933 541	141 566	102	17	6	5	1	1	1	12	1	395
15 108/-1	ICU	Y	MCME00000000	7108 153	199 276	92	31	18	4	5	3	1	17	13	446
57P31PA	COPD	Y	MCLY00000000	6486 503	426 852	36	46	23	5	11	7	1	12	7	274
13 121/-1	ICU	Y	MCMD00000000	6981 334	227 566	95	24	22	20	11	3	3	3	7	348
39 177	keratitis	Y	MCMC00000000	6682 085	464 516	53	25	6	5	1	7	4	6	7	449
KK1	CF	Y	MCMB00000000	6744 864	454 271	42	26	28	5	36	3	3	13	7	155
A5803	CAP	Y	MCMA00000000	6843 839	302 775	48	31	32	8	3	18	1	123	118	1567
TBCF10839	CF	Y	MCLZ00000000	6852 331	395 296	63	28	1	5	11	3	4	10	3	234

Scaff., number of scaffolds; CF, cystic fibrosis; Non-CF, non-CF clinical isolate; Non-CF Br, non-CF bronchiectasis; ICU, isolated from a patient in an intensive care unit; HE, hospital environment; tob. plant, tobacco plant; COPD, chronic obstructive pulmonary disease; CAP, community acquired pneumonia. NK, not known. MLST, multilocus sequence type, with individual allele numbers shown for the genes *acsA* (acs), *aroE* (aro), *guaA* (gua), *mutL* (mut), *nuoD* (nuo), *ppsA* (pps) and *trpE* (trp).

(bwa) tool (v0.7.5a; bwa-mem) (Li and Durbin 2009) with standard parameters. The reference genome (fasta) was first indexed with bwa index (Li and Durbin 2009) and samtools (Li *et al.* 2009) faidx. A sequence dictionary was created using picard-tools (http://broadinstitute.github.io/picard/; v1.135) CreateSequenceDictionary. The resulting sequence alignment map (*sam*) file from read mapping with bwa-mem was converted to a binary alignment map (*bam*) file using picard-tools SortSam and duplicates were marked using picard-tools MarkDuplicates. Finally, a bam file index was created with picard-tools BuildBamIndex. The Genome Analysis Toolkit (GATK) (McKenna *et al.* 2010) (v3.4.) Realignor Target Creator was used to designate targets for indel realignment and indels were realigned with GATK IndelRealigner. Variants were called using GATK HaplotypeCaller (-ploidy 1, –emitRefConfidence, GVCF) to produce a variant call file (*vcf*) that was genotyped using GATK GenotypeGVCFs and filtered using vcf tools (Danecek *et al.* 2011) vcffilter basic filtering (DP > 9 and QUAL > 10). Variant annotation was performed using snpEff (v4.1) (Cingolani *et al.* 2012) with the default parameters for gatk output (eff -gatk) to the reference genome database for PAO1 (uid57945). In addition, we evaluated whether a gene has a larger deletion not reported due to lack of sequencing reads for GATK or absence of genomic context in vcf files when predicting impact. First bam files were indexed with samtools index and the reads were aligned to a specified region (in this case a gene matching the coordinates in the snpEff database) using samtools depth. The results were processed to get an approximate 'alignment' length from which larger deletions could be determined. Deletions smaller than 30 bp were checked by aligning the reference gene with blastn (v 2.2.27+) (Camacho *et al.* 2009) to the assembled genome.

Resistome analysis

AMR genes were identified in all genomes based on the Comprehensive Antibiotic Resistance Database (CARD) (McArthur *et al.* 2013). This was done using the command-line version of the Resistance Gene Identifier (RGI) software, version 3.0.1 (McArthur *et al.* 2013). This software is based on BLASTP searches against the CARD, with curated e-value cut-offs to determine the presence of AMR genes, plus additional variant analysis.

Regions of genome plasticity, genomic islands and prophages

To identify regions of genome plasticity (RGPs), groups of orthologous proteins were computed using OrthoFinder v0.4 (Emms and Kelly 2015), resulting in 8819 orthogroups, out of which 1211 contained singletons. For draft genomes, contigs were reordered by similarity to a reference genome, as stated in Table S1 (Supporting Information), using IslandViewer 3 (Dhillon *et al.* 2015) to obtain a pseudochromosome. For each genome/pseudochromosome, an RGP was defined as a genomic region with at least two consecutive predicted coding sequences (CDS) conserved in 36 genomes compared or less. One conserved gene was allowed if surrounded by other CDS fulfilling the criteria, since transposable elements, often present in multiple copies and conserved across the strains, may otherwise be incorrectly split larger regions into smaller segments. The conserved CDS upstream and downstream of each RGP serving as genomic anchors and possible insertion sites were retrieved and their orthogroup was used to identify hotspots of RGPs along the PAO1 genome. Nucleotide sequence similarity between RGPs was scored using Mash (Ondov *et al.* 2016) and RGPs closer than a Mash distance of 0.04 were used to reconstruct groups of similar RGPs. Additional manual curation was performed in Cytoscape v3.4.0 (Shannon *et al.* 2003) to remove edges linking larger interconnected groups and a between-edge clustering was performed in R v3.3.3. To validate our findings, RGPs were compared to a manually curated dataset based on previous analyses and literature review for PAO1 (Mathee *et al.* 2008). GIs (clusters of genes of probable horizontal origin usually identified with cut-offs larger than for RGPs) were predicted using the comparative genomics approach of IslandPick (Langille, Hsiao and Brinkman 2008), plus the sequence composition-based approaches SIGI-HMM (Waack *et al.* 2006) and IslandPath-DIMOB v1.0.0 (Bertelli and Brinkman 2018), as available in IslandViewer 4 (Bertelli *et al.* 2017). Prophages were predicted using PHASTER (Arndt *et al.* 2016). All RGPs were further classified as GIs or prophage when overlapping their respective predictions. Further data processing was performed in R using packages GenomicRanges, igraph, plotrix, ape, phangorn, and vegan. The circular plot was produced using CIRCOS (Krzywinski *et al.* 2009).

RESULTS AND DISCUSSION

Distribution of the panel strain genomes amongst the wider *P. aeruginosa* population

Using core genome single nucleotide variant (SNV) phylogeny analysis of the panel strains alongside genome sequence data from strains publicly available on NCBI, we were able to place the panel strains in the wider context of the *P. aeruginosa* population (Fig. 1). The panel strains were widely distributed, with 31 strains in group 1 and 9 strains in group 2 (Fig. 1 and Table 1).

Loss-of-function mutations in panel strain genomes

The panel strain genomes were analysed for the presence of likely loss-of-function mutations that may be associated with known phenotypes. In particular, we focused on mutations that have been linked to adaptation during chronic infections of CF patients (summarised in Table 2). Several panel strains contain putative loss-of-function mutations in the gene encoding the virulence-related quorum-sensing regulator LasR, reported as a common adaptation in CF. They include five CF isolates, including representatives of four transmissible strains (LES400, AMT0023–34, AUS23, AUS52, KK1 and DK2). However, severe *lasR* mutations were also identified in the community acquired pneumonia isolate A5803, the burn-related isolate Mi162 and the tobacco plant isolate CPHL9433, indicating that such mutations are not restricted to CF. In a previous study (Cullen *et al.* 2015), these isolates were tested for pyocyanin production. Whilst the strains LES400, AMT0023–34, AUS23, AUS52, KK1, DK2 and Mi162 were amongst the low producers of pyocyanin, despite its *lasR* mutation strain CPHL9433 was one of the higher producers. Interestingly, strain CPHL9433 has a mutation in *gacA*, encoding part of the GacAS two-component regulatory system known to play a role in regulation of quorum sensing. It has been reported that *gacA* knockout mutants are impaired in their ability to produce pyocyanin (Kay *et al.* 2006). Hence, this strain is able to overcome two mutations predicted to lead to loss of this phenotype. Other low pyocyanin producers, such as C3719, AA43, AA44, 968333S, NH57388A, did not have clear *lasR* loss-of-function mutations. In strain 968333S, there is a mutation that would lead to a single amino acid change in LasR ($M_{212} \rightarrow R$). An analysis of other quorum-sensing—related genes (*las*, *rhl* and *pqs* genes) was conducted to look for other mutations that might explain this

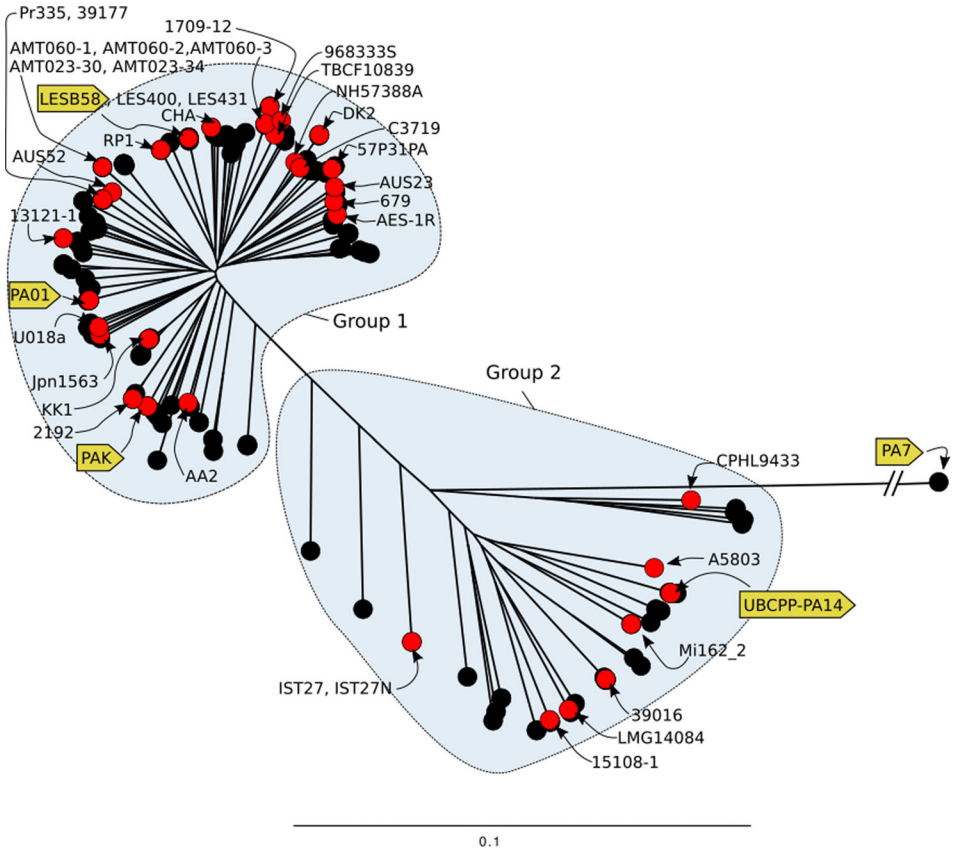

Figure 1. Core genome phylogeny based on 218 520 SNVs. Red dots identify panel strains, while black dots identify strains from NCBI. Commonly studied reference strains are identified by yellow boxes. The two main groups that define the population structure of *P. aeruginosa* are highlighted in light blue. Strain PA7, which clusters separately from these two groups (and is not in the panel), was included for comparison.

phenotype. In strain C3719, there is a 184 bp deletion in the *rhlI* gene. However, mutations in the targeted genes were not found in the other low pyocyanin producers.

Loss-of-function mutations in the genes encoding the component part of the MexAB-OprM efflux pump are also common in CF and bronchiectasis (Winstanley, O'Brien and Brockhurst 2016; Hilliam *et al.* 2017). Such mutations were found in the genomes of 12 of the panel isolates, all associated with CF infections. The genomes of the sequential CF isolates AA2, AA43 and AA44 all contain the same frameshift mutation in *mexA*. The related strains IST27 (mucoid) and IST27N (non-mucoid) contain the same frameshift mutation in *oprM*. The late CF isolate AMT0023–34 contains a premature stop codon in *mexB* not seen in the related early CF isolate AMT0023–30. The DK2 isolate has a 78 bp deletion in *mexB* and a frameshift in *mexA*. *mexB* mutations were also detected in the genomes of AUS23 and LES431, whilst a *mexA* mutation was also detected in the genome of NH57388A.

Another commonly reported CF adaptation is the occurrence of mucoid colonies, usually due to *mucA* mutations leading to overproduction of alginate. We found that nine of the panel isolates carry putative loss-of-function mutations in *mucA*. Eight of these isolates were isolated from CF patients. The ninth was 968333S, an isolate from a patient with non-CF bronchiectasis. Of the four strains included in this study and reported previously as producing the highest levels of alginate (AMT0060–2, CHA, IST27, 968333S) (Cullen *et al.* 2015), three carry putative *mucA* loss-of-function mutations (AMT0060–2, IST27 and 968333S; Table 2). In the fourth, strain CHA, there is a mutation leading to a single amino acid change (Sall *et al.* 2014). The presence of a *mucA* mutation in the genome does not guarantee that an isolate will have the mucoid phenotype because compensatory mutations can occur, leading to reversion to non-mucoid. IST27N is a spontaneous non-mucoid variant of the mucoid strain IST27 (De Soyza *et al.* 2013). However, we were unable to detect a compensatory mutation that could explain this reversion. It is clear that not all such mutations have been characterised.

The GacA/GacS two-component regulatory system has been implicated in the switch between acute and chronic infection lifestyles and plays a key role in virulence. Our analysis confirmed the presence of the previously reported *gacS* loss-of-function deletion mutations in the genome of CHA (Sall *et al.* 2014). We also identified frameshift mutations in the *gacS* genes of strain CPHL9433 (isolated from a tobacco plant) and the related CF isolates AMT0060–2, AMT0060–30 and AMT0060–34.

The analysis confirmed that strain 968333S, a known hypermutator, has an 11 bp frameshifting deletion in the *mutS* gene, but no other panel strains had putative loss-of-function mutations in any of the DNA mismatch repair genes, *mutS*, *mutL*, *mutM* and *uvrD*. Four isolate genomes contain a nonsense mutation in biofilm dispersal gene *rbdA*. They were isolated from CF (C3719, TBCF10839), the hospital environment (Pr335) and a (keratitis) eye infection (39177).

There were some mutations in genes associated with motility. As reported previously (Jeukens *et al.* 2014), the genomes of strains LES400 and LES431 have acquired a premature stop codon in *fleR*, implicated in loss of motility. We further observed that the non-motile isolate AUS23 has a frameshift mutation in the *fliG* gene. However, we could not identify any candidate

Table 2. Summary of loss-of-function mutations.

Isolate	Mutation	Isolate details
lasR mutants		
LES400	7 bp frameshift	CF (transmissible strain)
AUS23	Premature stop codon	CF (transmissible strain)
AUS52	Premature stop codon	CF (transmissible strain)
DK2	Gene deleted	CF (transmissible strain)
AMT0023–34	1 bp frameshift	CF (late isolate)
KK1	Gene deleted	CF
A5803	Premature stop codon	Community acquired pneumonia
Mi162	168 bp frameshift	Burn patient
CPHL9433	2 bp frameshift	Tobacco plant
mucA mutants		
AUS23	5 bp frameshift	CF (transmissible strain)
AUS52	Premature stop codon	CF (transmissible strain)
DK2	1 bp frameshift	CF (transmissible strain)
AMT0060–1	1 bp frameshift 1 bp frameshift	CF (late isolate)
AMT0060–2	1 bp frameshift	CF (late isolate)
NH57388A	89 bp deletion	CF
IS27 & IS27N	1 bp frameshift	CF
968333S	7 bp frameshift	Non-CF bronchiectasis
mexA-mexB-oprM mutants		
LES431	1 bp frameshift (*mexB*)	CF (transmissible strain)
AUS23	Premature stop codon (*mexB*)	CF (transmissible strain)
AUS52	1 bp frameshift (*mexA*)	CF (transmissible strain)
DK2	2 bp frameshift (*mexA*) 78 bp deletion (*mexB*)	CF (transmissible strain)
AMT0023–34	Premature stop codon (*mexB*)	CF (late isolate)
AA2	1 bp frameshift (*mexA*)	CF
NH57388A	1 bp frameshift (*mexA*)	CF
IS27 & IS27N	2 bp frameshift (*oprM*)	CF
mutS mutants		
968333S	11 bp frameshift	Non-CF bronchiectasis
gacAS mutants		
AMT0060–2	2 bp frameshift (*gacA*)	CF (late isolate)
AMT0023–30	2 bp frameshift (*gacA*)	CF (early isolate)
AMT0023–34	2 bp frameshift (*gacA*)	CF (late isolate)
CPHL9433	37 bp frameshift (*gacA*)	Tobacco plant
CHA	148 bp deletion (*gacS*)	CF
motility mutants		
LES400 & LES431	Premature stop codon (*fleR*)	CF (transmissible strain)
AUS23	1 bp frameshift (*fliG*)	CF (transmissible strain)

loss-of-function mutation in the genome of 968333S, also reported to be non-motile.

RGP in the panel strain genomes

Taking advantage of the phylogenetic distribution, and number, of genomes in the panel, the accessory genome of *P. aeruginosa* was characterised using comparative genomics approaches. A total of 2315 RGPs (regions containing at least two consecutive predicted genes that were absent from at least 10% of the genomes) were identified (Table S1, Supporting Information). All but four (25/29) of the curated regions of PAO1 larger than 2 kb were recovered with good congruence in RGP boundary definition, validating the method (Fig. 2). The three missed (and one poorly predicted) curated regions had been identified by pairwise comparison to various strains and are conserved in over 36 of the strains studied here, thereby likely representing regions of lesser plasticity. For example, one curated region had been identified by comparison to PA7, a more distantly related strain absent from the panel genomes (Roy *et al.* 2010; Klockgether *et al.* 2011).

The clustering of RGPs by sequence similarity reveals that most regions are found uniquely in a few strains (Fig. 3A). This is likely due primarily to the high diversity of *P. aeruginosa* genomes and suggests that this genus must be sampled further to better characterise the diversity of some *P. aeruginosa* lineages. To a lesser extent, incomplete genome sequencing likely impacts RGP definition and clustering, as small contigs are not always accurately placed. As previously observed (Klockgether *et al.* 2011), RGPs are scattered around the genome (Fig. 2). GI and prophage predictions overlap respectively with 43% and 16% of the RGPs encoding more than four genes, suggesting that these regions

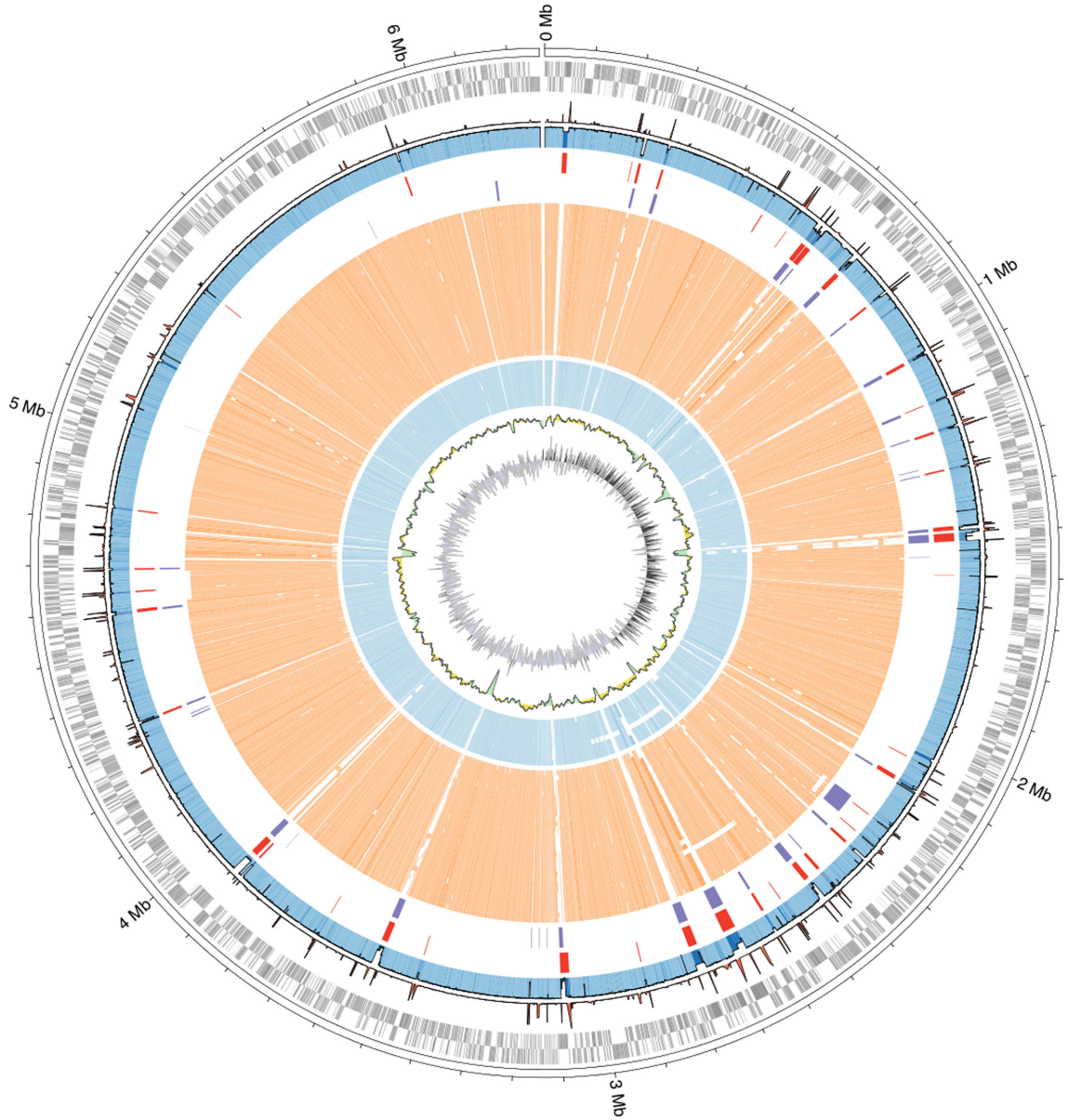

Figure 2. Circular genome view, illustrating the distribution and conservation of predicted RGPs using the *P. aeruginosa* PAO1 genome. From the outer to the inner circle: genes on the plus and minus strands (grey), the number of RGPs in the 40 strains bordered by conserved genes based on the orthogroups of proteins (red peaks), the number of orthologs of PAO1 proteins (blue), the predicted RGPs (dark red), curated literature RGPs (purple) and the presence of orthologs of PAO1 proteins in the 39 other *Pseudomonas* panel genomes belonging to group 1 (orange) and group 2 (light blue), GC content (green/yellow) and GC skew (purple).

have been acquired horizontally (Fig. 3B). Most of the RGPs, including GIs and prophages, previously described (Winstanley *et al.* 2009; Klockgether *et al.* 2011) were identified in the reference genomes of *P. aeruginosa* PAO1, PA14 and LESB58 (Table S1, Supporting Information).

Hierarchical clustering and neighbour-joining reconstructions, based on the presence/absence of each group of RGP in the panel strains (Fig. 3B and C), clearly separates the two major groups of *P. aeruginosa* shown in Fig. 1, and successfully groups very close monophyletic strains. Nevertheless, the Robinson-Foulds distance between the core genome SNV phylogeny and RGP presence-absence phylogenies is high (42–48). Thus, although the presence/absence of groups of RGPs lacks resolution, it still harbours some phylogenetic signal. This suggests that, at least at the largest clade grouping level, there may be distinct accessory regions, GI and prophage gene pools that each large clade is drawing upon. The analysis of additional genomes could improve the resolution of the tree and further reveal the association of different mobile gene pools with different clades.

In addition to the RGPs, we identified the presence of two very large deletions with distinct boundaries (2 950 111 to 3 129 523 and 2 972 067 to 3 174 547 of PA14) in strains AMT0023–34 and Mi162˙2 isolated from CF and burn patients, respectively. A similar event with no mention of a mobile element in this region had previously been observed in a CF isolate RN43 with no apparent growth defect (Cramer *et al.* 2011). Our findings in two strains belonging to the two major groups (Fig. 1) suggest that this 179 kb genomic region close to the terminus of replication (around 3.219 Mb in PA14) is dispensable and prone to deletion in *P. aeruginosa* strains.

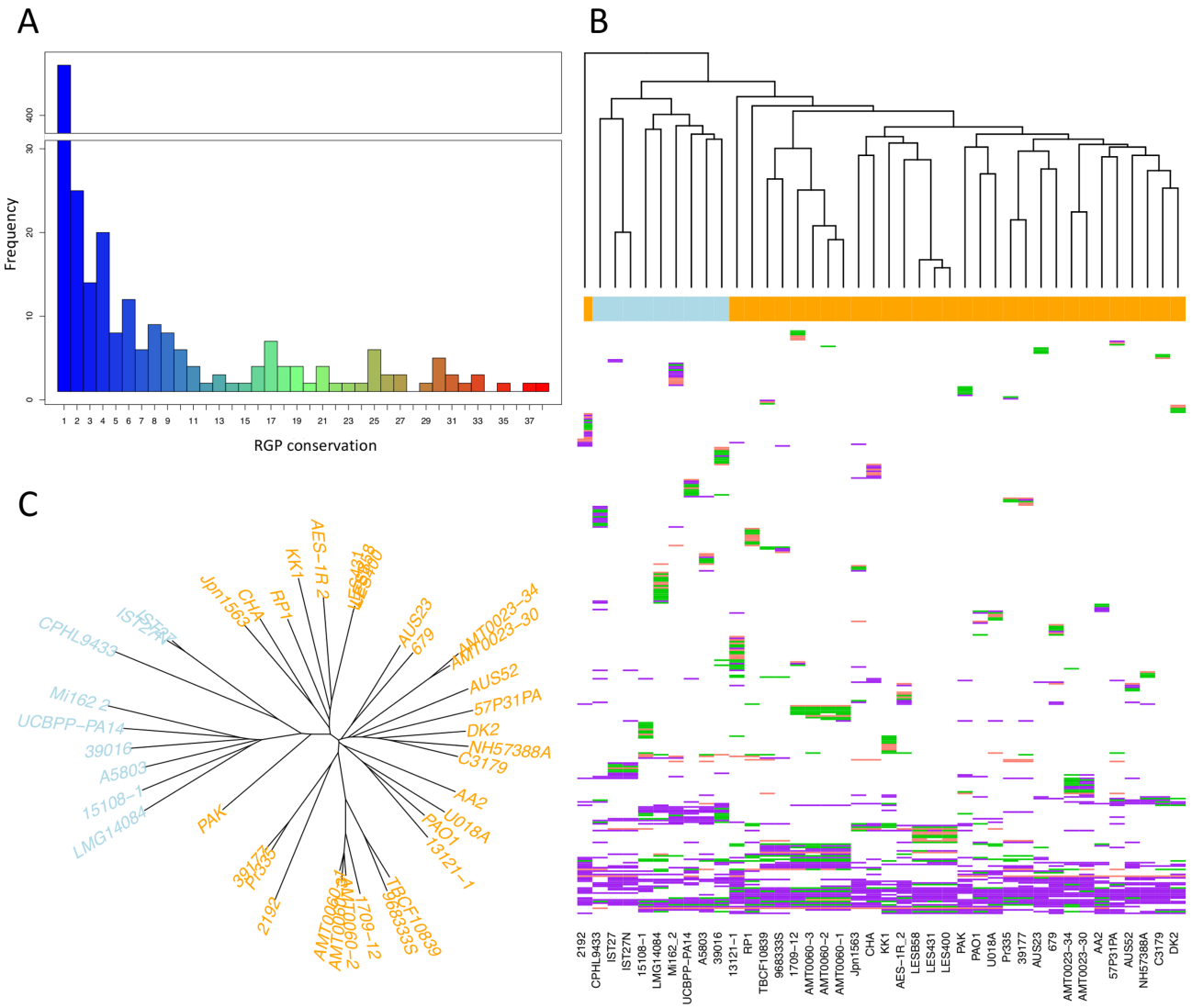

Figure 3. Conservation of genomic islands (GIs), phages and other regions of genome plasticity (RGPs). (**A**) Conservation of RGPs among the 40 genomes, showing that most regions are found uniquely in a few strains. (**B**) Hierarchical clustering of strains based on the presence-absence of RGPs. The two main groups of *P. aeruginosa* strains are indicated below the cladogram in orange (group 1) and blue (group 2). The prediction of RGPs as probable phages or other GIs is indicated in green and salmon, respectively. Other RGPs are shown in purple. (**C**) Neighbour-joining phylogenetic tree, based on a distance matrix of the percentage of shared RGP groups, clusters the main groups of *P. aeruginosa* (group 1; orange, group 2; blue labels) similarly to that of the core genome phylogeny shown in Fig. 1. The outgroup PA7 shown in the tree of Fig. 1 branches close to strain CPHL9433.

AMR genes and mutations in the panel strain genomes

We characterised the resistome of the panel strains using a database approach (Fig. 4). From the 40 genomes analysed, 37 unique resistomes were identified, thus reinforcing the considerable diversity observed in antibiotic susceptibility for these strains (Cullen *et al.* 2015). However, the observation that CF strains generally showed resistance to more antibiotics than non-CF strains was not as clear when looking at the resistome data. In fact, attempting to relate these results with previously determined antimicrobial susceptibility data (Cullen *et al.* 2015) was difficult. This is likely to be due to the non-specific nature and expression level dependence of efflux mechanisms (Blair *et al.* 2015). Only resistance to quinolones (Nakano *et al.* 1997; Lee *et al.* 2005) was relatively easy to associate with specific *gyr* variants. This difficulty has been highlighted previously for *P. aeruginosa*. Jeukens *et al.* (2017) have demonstrated this by focusing on a limited set of strains, including LESB58, which is on the more 'resistant' side of the panel, and PAO1, on the 'susceptible' side. Expression levels of the intrinsic gene *ampC* appeared more likely to underlie differences in beta-lactam resistance (Cabot *et al.* 2011) than the variant of *Pseudomonas*-derived cephalosporinase or AmpC beta-lactamase present. In addition, differences in the resistance to aminoglycosides have been attributed mostly to the regulation of efflux mechanisms (Poole 2005; Garneau-Tsodikova and Labby 2016). The *gyr* variant found in LESB58 could reasonably account for ciprofloxacin and levofloxacin (quinolones) resistance, yet it does not account for quinolone resistance in LES400, for instance. Efflux pumps do also have an impact on quinolone resistance in *P. aeruginosa* (Jalal *et al.* 2000; Lomovskaya *et al.* 2001; Kriengkauykiat *et al.* 2005).

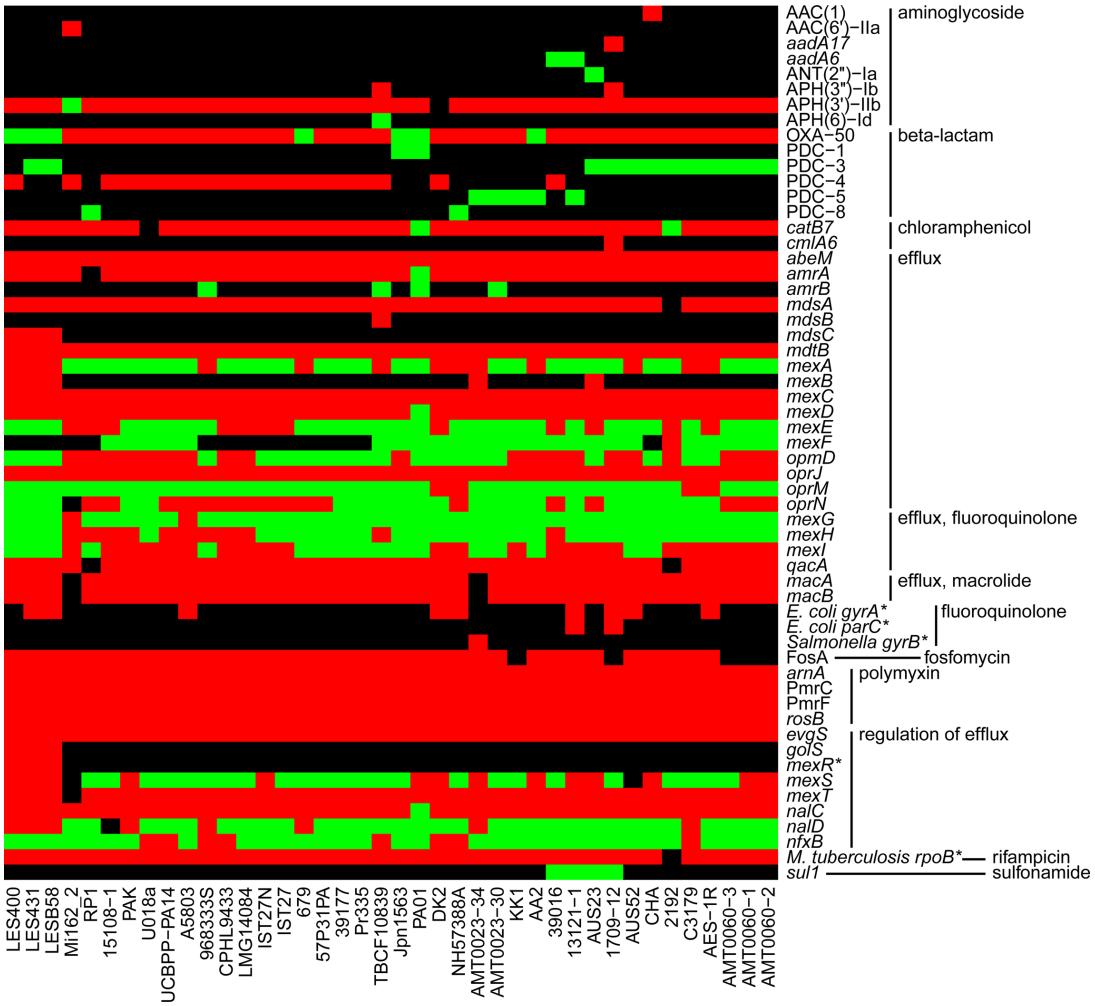

Figure 4. Resistome of the panel strains. Gene or variant (*) presence was determined using the RGI-CARD (McArthur *et al.* 2013). AMR genes are grouped by antibiotic family or function. Green: perfect match to a gene or variant (*) in the CARD, red: similar to a gene in the CARD, according to curated cut-offs, black: no match in the CARD. Genomes are ordered based on hierarchical clustering of the resistomes (dendrogram not shown).

CONCLUSIONS

We have demonstrated that the reference panel of isolates harbours substantial phylogenetic diversity, and includes representatives in both of the major *P. aeruginosa* groups (groups 1 and 2). It was possible to identify loss-of-function mutations indicative of adaptation, especially amongst isolates associated with chronic respiratory infections, but our study further demonstrates the difficulty in relating genomics data to *P. aeruginosa* isolate phenotypes, especially in relation to AMR. These difficulties reflect both the diversity of the strains included in the panel and the complexity of the regulatory networks that control virulence and other functions in *P. aeruginosa* (Balasubramanian *et al.* 2013). Much of our knowledge to date has relied on close analysis of a limited number of laboratory reference strains. Our findings demonstrate the need to extend beyond this to capture the diversity of the species. Our examination of the accessory genome content indicated that group 1 and group 2 isolates also form separate clusters based on mobile gene content. The analysis of additional genomes in these diverse genera could improve the resolution of the tree and further reveal the degree of association of different mobile gene pools with different clades/taxonomic levels, including genes of medical interest, such as those associated with AMR.

SUPPLEMENTARY DATA

Supplementary data are available at *FEMSLE* online.

ACKNOWLEDGEMENTS

The following authors (ADS, SMcC and CW) were all members of the EU COST Action BM1003: Microbial cell surface determinants of virulence as targets for new therapeutics in cystic fibrosis (http://www.cost-bm1003.info/) and acknowledge this support in the collation of the strain panel.

FUNDING

This work was supported by Cystic Fibrosis Canada [RCL], Genome Canada [FSLB], a Swiss National Science Foundation fellowship [P300PA˙164673 to CB], Société Académique Vaudoise [CB], Deutsche Forschungsgmeinschaft [SFB900, project A2 to JK], Action Medical Research [GN2444 to JLF], Medical Research Foundation [MRF-091–0006-RG-FOTHE to JLF] and the UK Cystic Fibrosis Trust [RS34 to CW].

Conflict of interest. None declared.

REFERENCES

Arndt D, Grant JR, Marcu A *et al*. PHASTER: a better, faster version of the PHAST phage search tool. *Nucleic Acids Res* 2016;**44**:W16–21.

Balasubramanian D, Schneper L, Kumari H *et al*. A dynamic and intricate regulatory network determines *Pseudomonas aeruginosa* virulence. *Nucleic Acids Res* 2013;**41**:1–20.

Bertelli C, Brinkman FSL. Improved genomic island predictions with IslandPath - DIMOB. *Bioinformatics* 2018;**34**:bty095.

Bertelli C, Laird MR, Williams KP *et al*. IslandViewer 4: expanded prediction of genomic islands for larger-scale datasets. *Nucleic Acids Res* 2017;**45**:W30–W35.

Blair JM, Webber MA, Baylay AJ *et al*. Molecular mechanisms of antibiotic resistance. *Nat Rev Microbiol* 2015;**13**:42–51.

Cabot G, Ocampo-Sosa AA, Tubau F *et al*. Overexpression of AmpC and efflux pumps in *Pseudomonas aeruginosa* isolates from bloodstream infections: prevalence and impact on resistance in a Spanish multicenter study. *Antimicrob Agents Ch* 2011;**55**:1906–11.

Camacho C, Coulouris G, Avagyan V *et al*. BLAST+: architecture and applications. *BMC Bioinformatics* 2009;**10**:421.

Cingolani P, Platts A, Wang le L *et al*. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. *Fly* 2012;**6**:80–92.

Cohen TS, Prince A. Cystic fibrosis: A mucosal immunodeficiency syndrome. *Nat Med* 2012;**18**:509–19.

Cramer N, Klockgether J, Wrasman K *et al*. Microevolution of the major common *Pseudomonas aeruginosa* clones C and PA14 in cystic fibrosis lungs. *Environ Microbiol* 2011;**13**:1690–704.

Cramer N, Wiehlmann L, Ciofu O *et al*. Molecular epidemiology of chronic *Pseudomonas aeruginosa* airway infections in cystic fibrosis. *PLoS One* 2012;**7**:e50731.

Cullen L, Weiser R, Olszak T *et al*. Phenotypic characterization of an international *Pseudomonas aeruginosa* reference panel: strains of cystic fibrosis (CF) origin show less in vivo virulence than non-CF strains. *Microbiology* 2015;**161**:1961–77.

Danecek P, Auton A, Abecasis G *et al*. The variant call format and VCFtools. *Bioinformatics* 2011;**27**:2156–8.

De Soyza A, Hall AJ, Mahenthiralingam E *et al*. Developing an international *Pseudomonas aeruginosa* reference panel. *MicrobiologyOpen* 2013;**2**:1010–23.

Dhillon BK, Laird MR, Shay JA *et al*. IslandViewer 3: more flexible, interactive genomic island discovery, visualization and analysis. *Nucleic Acids Res* 2015;**43**:W104–8.

Emms DM, Kelly S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol* 2015;**16**:157.

Fischer S, Klockgether J, Moran Losada P *et al*. Intraclonal genome diversity of the major *Pseudomonas aeruginosa* clones C and PA14. *Environ Microbiol Rep* 2016;**8**:227–34.

Freschi L, Jeukens J, Kukavica-Ibrulj I *et al*. Clinical utilization of genomics data produced by the international *Pseudomonas aeruginosa* consortium. *Front Microbiol* 2015;**6**:1036.

Garneau-Tsodikova S, Labby KJ. Mechanisms of resistance to aminoglycoside antibiotics: overview and perspectives. *Med Chem Commun* 2016;**7**:11–27.

Hilker R, Munder A, Klockgether J *et al*. Interclonal gradient of virulence in the *Pseudomonas aeruginosa* pangenome from disease and environment. *Environ Microbiol* 2015;**17**:29–46.

Hilliam Y, Moore MP, Lamont IL *et al*. *Pseudomonas aeruginosa* adaptation and diversification in the non-cystic fibrosis bronchiectasis lung. *Eur Respir J* 2017;**49**: 1602108.

Jalal S, Ciofu O, Hoiby N *et al*. Molecular mechanisms of fluoroquinolone resistance in *Pseudomonas aeruginosa* isolates from cystic fibrosis patients. *Antimicrob Agents Ch* 2000;**44**: 710–2.

Jeukens J, Boyle B, Kukavica-Ibrulj I *et al*. Comparative genomics of isolates of a *Pseudomonas aeruginosa* epidemic strain associated with chronic lung infections of cystic fibrosis patients. *PLoS One* 2014;**9**:e87611.

Jeukens J, Kukavica-Ibrulj I, Emond-Rheault JG *et al*. Comparative genomics of a drug-resistant *Pseudomonas aeruginosa* panel and the challenges of antimicrobial resistance prediction from genomes. *FEMS Microbiol Lett* 2017;**364**, DOI: 10.1093/femsle/fnx161.

Kay E, Humair B, Denervaud V *et al*. Two GacA-dependent small RNAs modulate the quorum-sensing response in *Pseudomonas aeruginosa*. *J Bacteriol* 2006;**188**:6026–33.

Klockgether J, Cramer N, Wiehlmann L *et al*. *Pseudomonas aeruginosa* genomic structure and diversity. *Front Microbiol* 2011;**2**:150.

Klockgether J, Munder A, Neugebauer J *et al*. Genome diversity of *Pseudomonas aeruginosa* PAO1 laboratory strains. *J Bacteriol* 2010;**192**:1113–21.

Kos VN, Deraspe M, McLaughlin RE *et al*. The resistome of *Pseudomonas aeruginosa* in relationship to phenotypic susceptibility. *Antimicrob Agents Ch* 2015;**59**:427–36.

Kriengkauykiat J, Porter E, Lomovskaya O *et al*. Use of an efflux pump inhibitor to determine the prevalence of efflux pump-mediated fluoroquinolone resistance and multidrug resistance in *Pseudomonas aeruginosa*. *Antimicrob Agents Ch* 2005;**49**:565–70.

Krzywinski M, Schein J, Birol I *et al*. Circos: an information aesthetic for comparative genomics. *Genome Res* 2009;**19**:1639–45.

Langille MG, Hsiao WW, Brinkman FS. Evaluation of genomic island predictors using a comparative genomics approach. *BMC Bioinformatics* 2008;**9**:329.

Lee JK, Lee YS, Park YK *et al*. Alterations in the GyrA and GyrB subunits of topoisomerase II and the ParC and ParE subunits of topoisomerase IV in ciprofloxacin-resistant clinical isolates of *Pseudomonas aeruginosa*. *Int J Antimicrob Ag* 2005;**25**:290–5.

Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 2009;**25**:1754–60.

Li H, Handsaker B, Wysoker A *et al*. The sequence alignment/map format and SAMtools. *Bioinformatics* 2009;**25**:2078–9.

Lomovskaya O, Warren MS, Lee A *et al*. Identification and characterization of inhibitors of multidrug resistance efflux pumps in *Pseudomonas aeruginosa*: novel agents for combination therapy. *Antimicrob Agents Ch* 2001;**45**:105–16.

Lyczak JB, Cannon CL, Pier GB. Establishment of *Pseudomonas aeruginosa* infection: lessons from a versatile opportunist. *Microbes Infect* 2000;**2**:1051–60.

McArthur AG, Waglechner N, Nizam F *et al*. The comprehensive antibiotic resistance database. *Antimicrob Agents Ch* 2013;**57**:3348–57.

McKenna A, Hanna M, Banks E *et al*. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 2010;**20**:1297–303.

Mathee K, Narasimhan G, Valdes C *et al*. Dynamics of *Pseudomonas aeruginosa* genome evolution. *Proc Natl Acad Sci USA* 2008;**105**:3100–5.

Nakano M, Deguchi T, Kawamura T *et al*. Mutations in the gyrA and parC genes in fluoroquinolone-resistant

clinical isolates of *Pseudomonas aeruginosa*. *Antimicrob Agents Ch* 1997;**41**:2289–91.

Nathwani D, Raman G, Sulham K *et al*. Clinical and economic consequences of hospital-acquired resistant and multidrug-resistant *Pseudomonas aeruginosa* infections: a systematic review and meta-analysis. *Antimicrob Resist In* 2014;**3**:32.

Ondov BD, Treangen TJ, Melsted P *et al*. Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol* 2016;**17**:132.

Pendleton JN, Gorman SP, Gilmore BF. Clinical relevance of the ESKAPE pathogens. *Expert Rev Anti Infe* 2013;**11**:297–308.

Pohl S, Klockgether J, Eckweiler D *et al*. The extensive set of accessory *Pseudomonas aeruginosa* genomic components. *FEMS Microbiol Lett* 2014;**356**:235–41.

Poole K. Aminoglycoside resistance in *Pseudomonas aeruginosa*. *Antimicrob Agents Ch* 2005;**49**:479–87.

Roy PH, Tetu SG, Larouche A *et al*. Complete genome sequence of the multiresistant taxonomic outlier *Pseudomonas aeruginosa* PA7. *PLoS One* 2010;**5**:e8842.

Sall KM, Casabona MG, Bordi C *et al*. A gacS deletion in *Pseudomonas aeruginosa* cystic fibrosis isolate CHA shapes its virulence. *PLoS One* 2014;**9**:e95936.

Shannon P, Markiel A, Ozier O *et al*. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* 2003;**13**:2498–504.

Stewart L, Ford A, Sangal V *et al*. Draft genomes of 12 host-adapted and environmental isolates of *Pseudomonas aeruginosa* and their positions in the core genome phylogeny. *Pathog Dis* 2014;**71**:20–25.

Stover CK, Pham XQ, Erwin AL *et al*. Complete genome sequence of *Pseudomonas aeruginosa* PAO1, an opportunistic pathogen. *Nature* 2000;**406**:959–64.

Treangen TJ, Ondov BD, Koren S *et al*. The Harvest suite for rapid core-genome alignment and visualization of thousands of intraspecific microbial genomes. *Genome Biol* 2014;**15**:524.

Tritt A, Eisen JA, Facciotti MT *et al*. An integrated pipeline for de novo assembly of microbial genomes. *PLoS One* 2012;**7**:e42304.

van Belkum A, Soriaga LB, LaFave MC *et al*. Phylogenetic distribution of CRISPR-Cas systems in antibiotic-resistant *Pseudomonas aeruginosa*. *mBio* 2015;**6**:e01796–01715.

Waack S, Keller O, Asper R *et al*. Score-based prediction of genomic islands in prokaryotic genomes using hidden Markov models. *BMC Bioinformatics* 2006;**7**:142.

Williams D, Evans B, Haldenby S *et al*. Divergent, coexisting *Pseudomonas aeruginosa* lineages in chronic cystic fibrosis lung infections. *Am J Resp Crit Care* 2015;**191**:775–85.

Winstanley C, Langille MG, Fothergill JL *et al*. Newly introduced genomic prophage islands are critical determinants of in vivo competitiveness in the liverpool epidemic strain of *Pseudomonas aeruginosa*. *Genome Res* 2009;**19**:12–23.

Winstanley C, O'Brien S, Brockhurst MA. *Pseudomonas aeruginosa* evolutionary adaptation and diversification in cystic fibrosis chronic lung infections. *Trends Microbiol* 2016;**24**:327–37.